

Introduction

- Set the scale:** measure storage metrics from running experiments to set the scale on expected bandwidth, typical file size, number of clients, etc.
 - http://home.fnal.gov/~garzogli/storage/dzoro-sam-file-access.html
 - http://home.fnal.gov/~garzogli/storage/cdf-sam-file-access-per-app-family.html
- Install storage solutions on FermiCloud testbed:** Lustre, BlueArc, Hadoop, OrangeFS
- Measure performance**
 - Run standard benchmarks on storage installations.
 - Study response of the technology to real-life (skim) applications access patterns (root-based)
 - Use HEPiX storage group infrastructure to characterize response to Intensity Frontier (IF) applications
- Fault tolerance:** simulate faults and study reactions
- Operations:** comment on potential operational issues. Clients on Virtual Machines: can we take advantage of the flexibility of cloud resources?

Data Access Tests

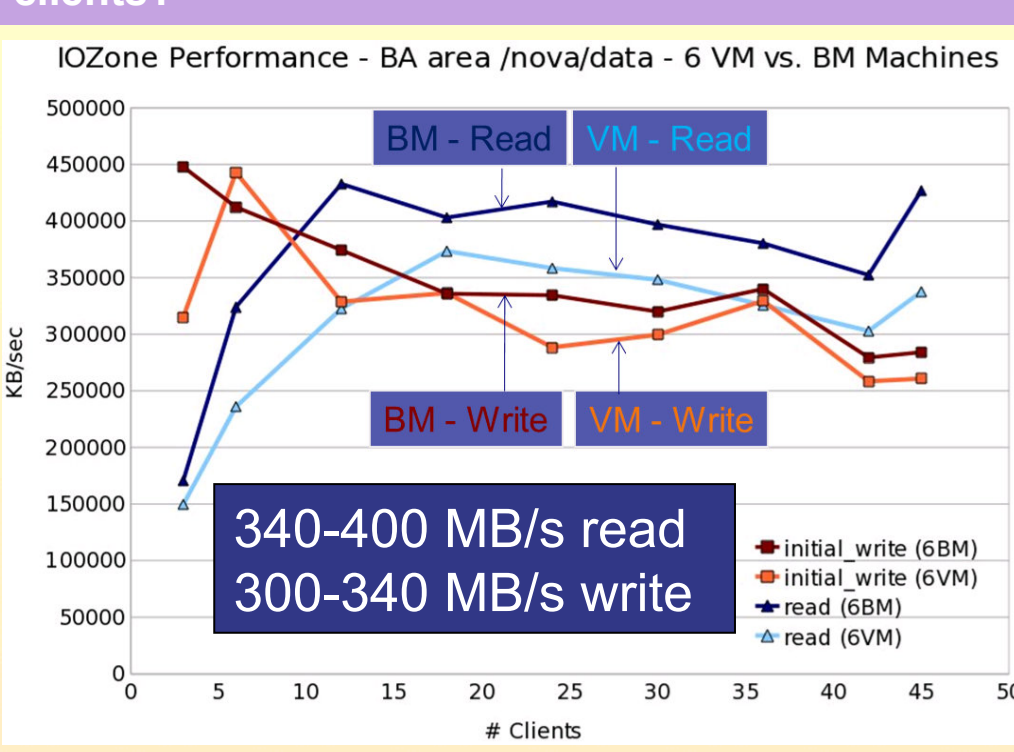
IOZone
Writes 2GB file from each client and performs read/write tests.
Setup: 3-60clients on Bare Metal (BM) and 3-21 VM/nodes.

Root-based applications
Used off-line root-based framework (ana) of the Nova neutrino Intensity Frontier (IF) experiment. Ran a "skim job" that read a data file and discarded large fraction of events. Reads stressed storage access; writes proved CPU-bound"
Setup: 3-60clients on Bare Metal and 3-21 VM/nodes.

MDTest
Different metadata FS operations on up to 50k files / dirs using different access patterns.
Setup: 21-504 clients on 21 VM.

BlueArc

How well do VM clients perform vs. Bare Metal clients?

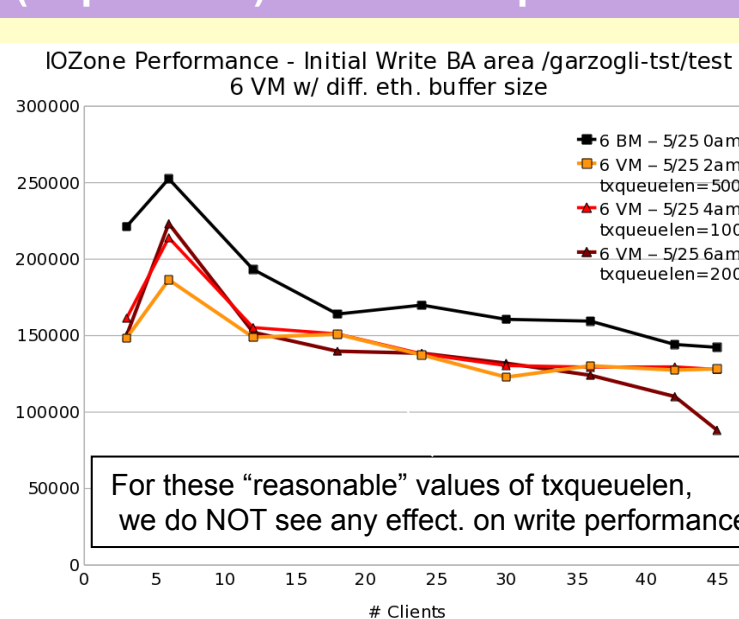


340-400 MB/s read
300-340 MB/s write

- Bare Metal Reads are (~10%) faster than VM Reads.
- Bare Metal Writes are (~5%) faster than VM Writes.

Note: results vary depending on the overall system conditions (net, storage, etc.)

Do transmit ethernet buffer sizes (txqueuelen) affect write performance?



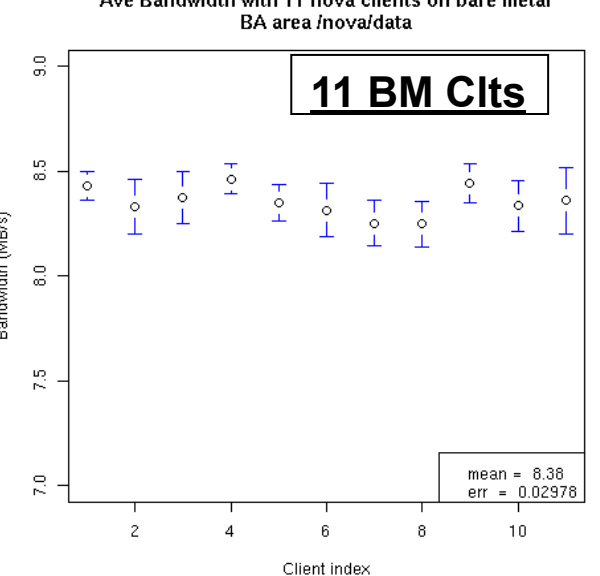
For these "reasonable" values of txqueuelen, we do NOT see any effect on write performance

Eth interface	txqueuelen
Host	1000
Host / VM bridge	500, 1000, 2000
VM	1000

Varying the eth buffer size for the client VMs does not change read / write BW.

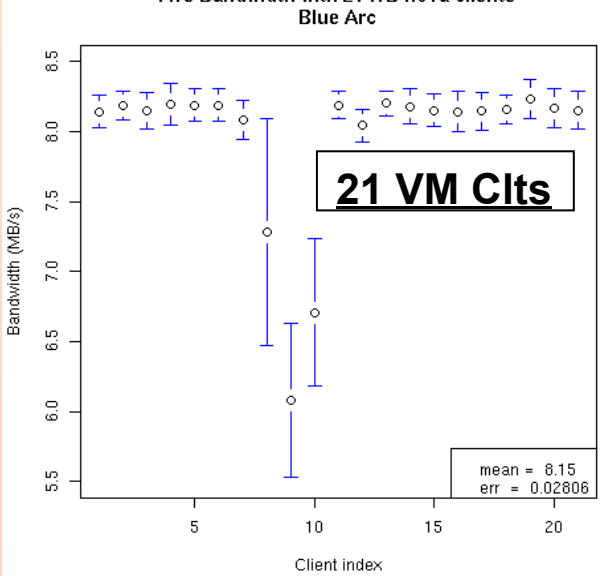
How well do VM clients perform vs. Bare Metal (BM) clients?

11 BM CIts



Root-app Read Rates:
21 CIts: 8.15 ± 0.03 MB/s
(Lustre: 12.55 ± 0.06 MB/s
Hadoop: ~7.9 ± 0.1 MB/s
OrangeFS: ~8.1 ± 0.1 MB/s)

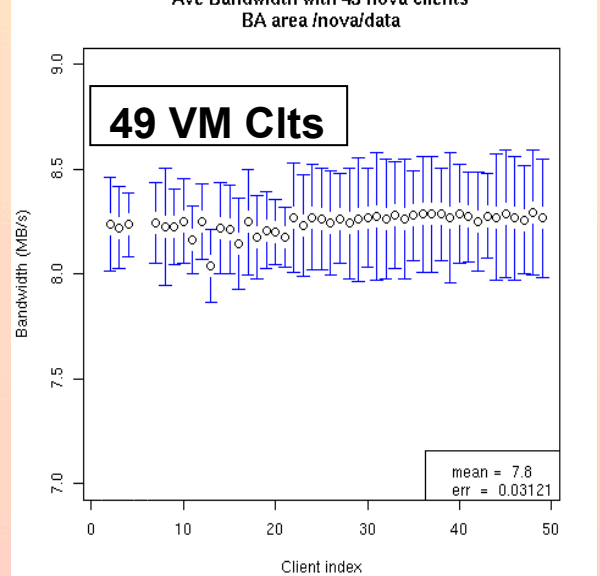
21 VM CIts



Read BW is essentially the same on Bare Metal and VM.

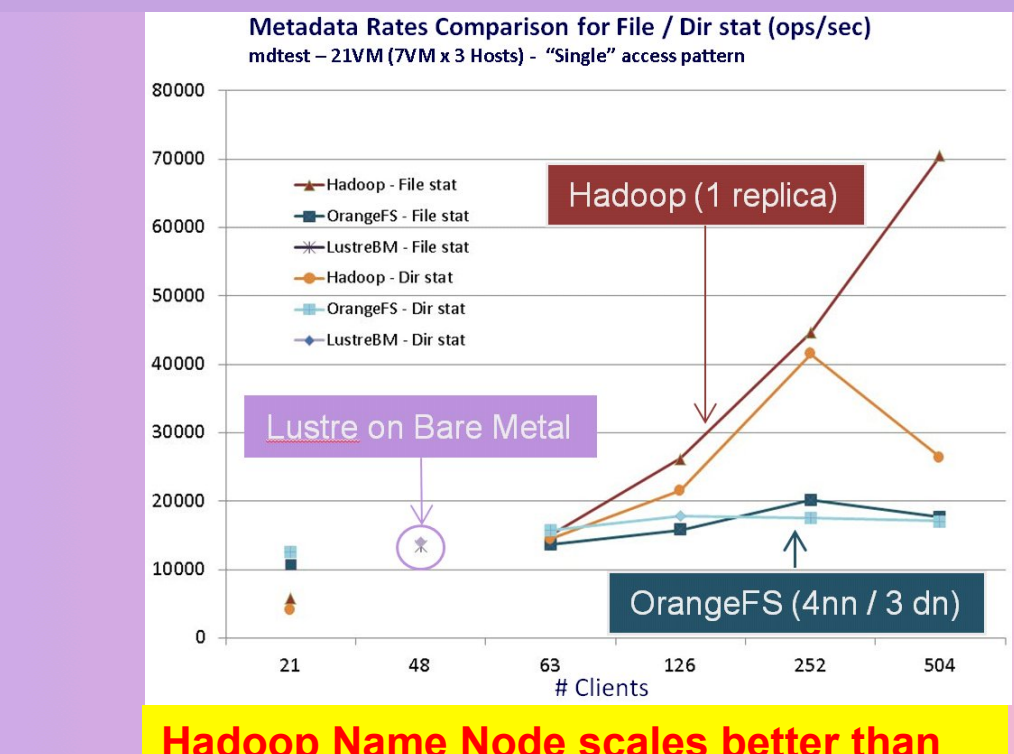
Note: NOVA skimming app reads 50% of the events by design. On BA, OrangeFS, and Hadoop, clients transfer 50% of the file. On Lustre 85%, because the default read-ahead configuration is inadequate for this use case.

49 VM CIts



MetaData Comparison

How well do name nodes scale with number of clients?



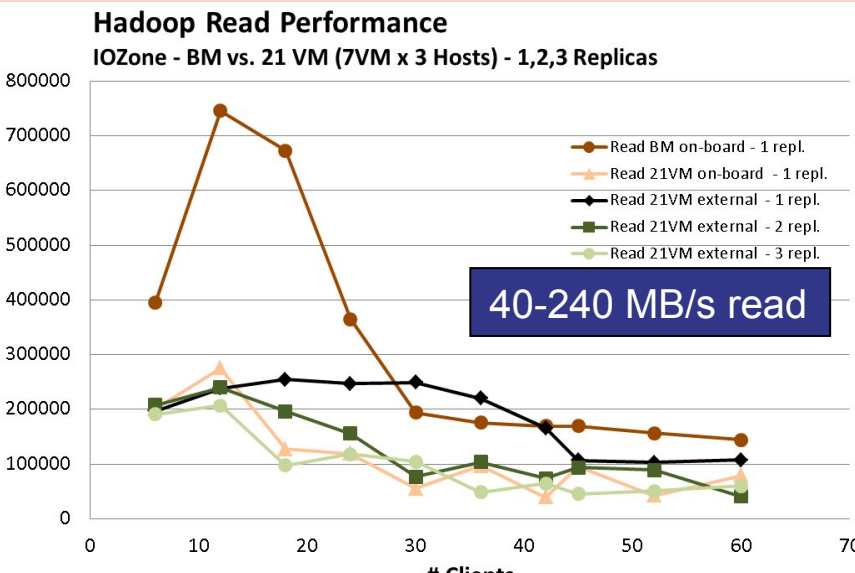
Hadoop Name Node scales better than OrangeFS

Hadoop

How well do VM clients perform vs. Bare Metal clients? Is there a difference for External vs. OnBoard clients? How does number of replica change performance?

Hadoop Read Performance

IOZone - BM vs. 21 VM (VPM x 3 Hosts) - 1,2,3 Replicas

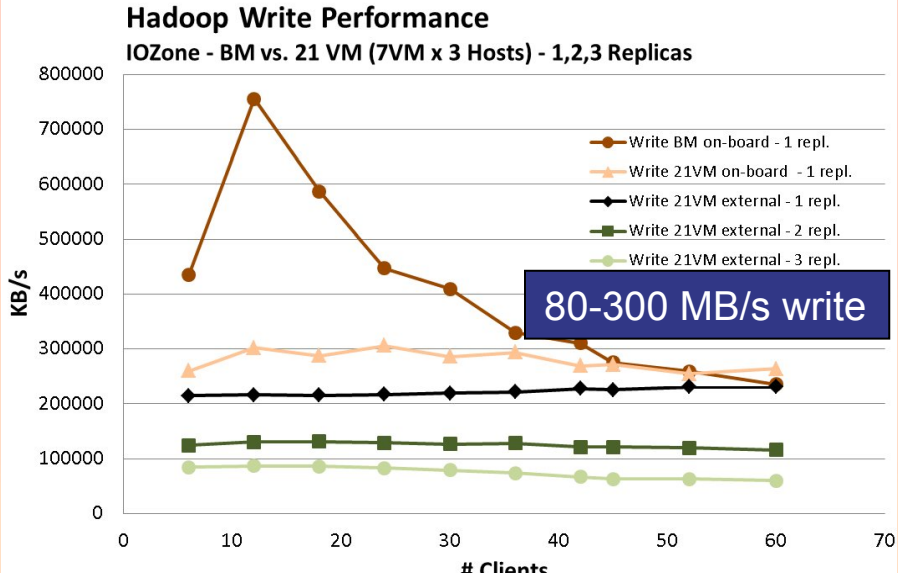


40-240 MB/s read

- On-Board Bare Metal clients reads gain from kernel caching, for a few clients. For many clients, same or ~50% faster than On-Board VM and ~100% faster than External VM clients.
- External VM Clients up to 100% faster than On-Board VM clients.
- Multiple replicas have little effect on read BW.

Hadoop Write Performance

IOZone - BM vs. 21 VM (VPM x 3 Hosts) - 1,2,3 Replicas

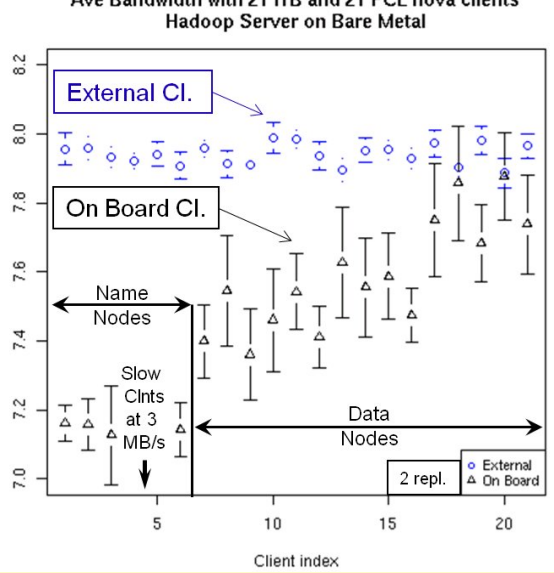


80-300 MB/s write

- On-Board Bare Metal client writes gain from kernel caching; generally faster than VM clients
- On-Board VM client 50%-200% faster than External VM clients.
- All VM write scale well with number of clients.
- For External VM clients, write speed scales almost linearly with the number of replicas.

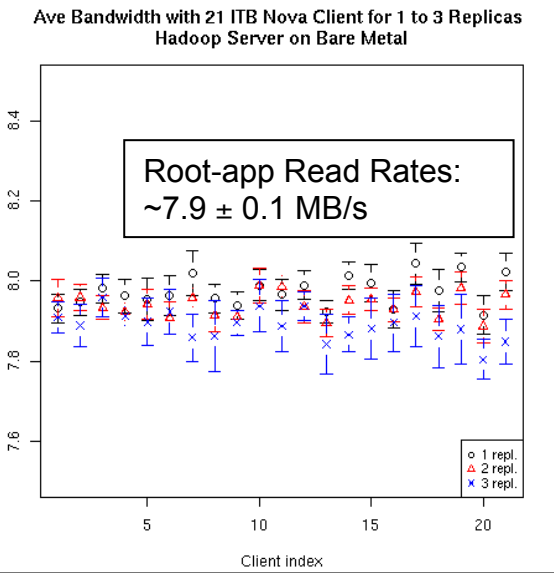
Root Benchmark

How does read BW vary for On-Brd vs. Ext. clnts?



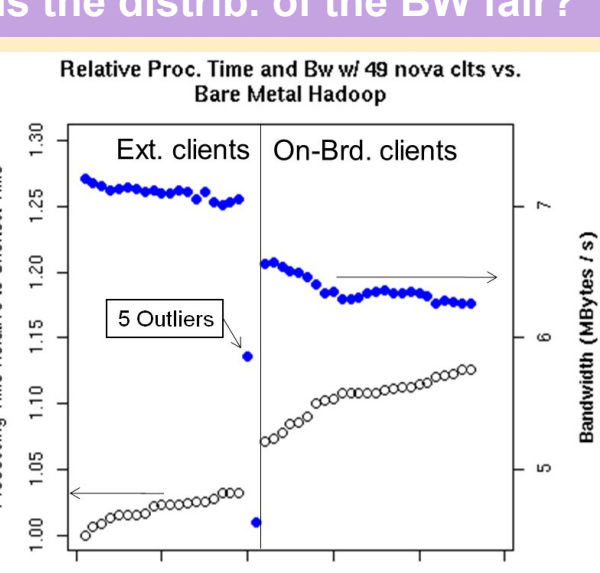
Ext (ITB) clients read ~5% faster than on-board (FCL) clients.

How does read bw vary vs. number of replicas?



Root-app Read Rates: ~7.9 ± 0.1 MB/s
(Lustre on Bare Metal was 12.55 ± 0.06 MB/s Read)
Number of replicas has minimal impact on read bandwidth.

49 clts (1 proc. / VM / core) saturate the BW to the srv. Is the distrib. of the BW fair?



At saturation, External clients read ~10% faster than On-Brd clients.. (Same as OrangeFS. Different from Lustre)
External and On-Board clients get the same share of the bw among themselves (within ~2%).

Conclusions

- Lustre on Bare Metal has the best performance as an external storage solution for the root skim application use case (fast read / little write). Consistent performance for general operations (tested via iozone)
- Consider operational drawback of special kernel
 - On-board clients only via virtualization, but server VM allows only slow write.
- Hadoop, OrangeFS, and BlueArc have equivalent performance for the root skim use case.
- Hadoop has good operational properties (maintenance, fault tolerance) and a fast name server, but performance is not impressive.
- BlueArc at FNAL is a good alternative for general operations since it is a well known production quality solution.
- The results of the study support the growing deployment of Lustre at Fermilab, while maintaining the BlueArc infrastructure.

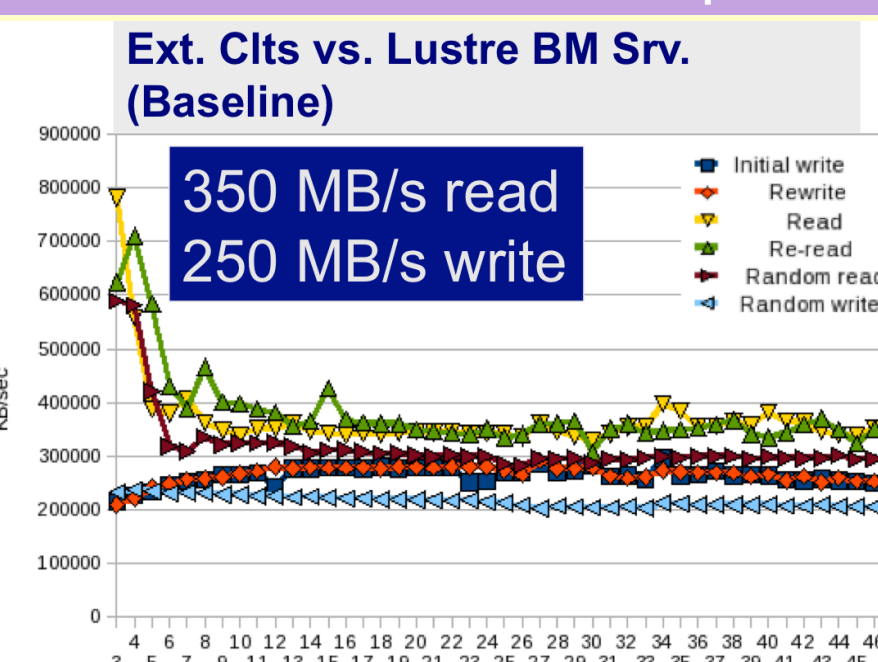
Storage	Benchmark	Read (MB/s)	Write (MB/s)	Notes
Lustre	IOZone	350	250 (70 on VM)	
	Root-based	12.6	-	
Hadoop	IOZone	50 - 240	80 - 300	Varies on replicas
	Root-based	7.9	-	
BlueArc	IOZone	340 - 400	300 - 340	Varies on conditions
	Root-based	8.2	-	
OrangeFS	IOZone	150-330	220-350	Varies on name nodes
	Root-based	8.1	-	

Lustre

Lustre Srv on VM

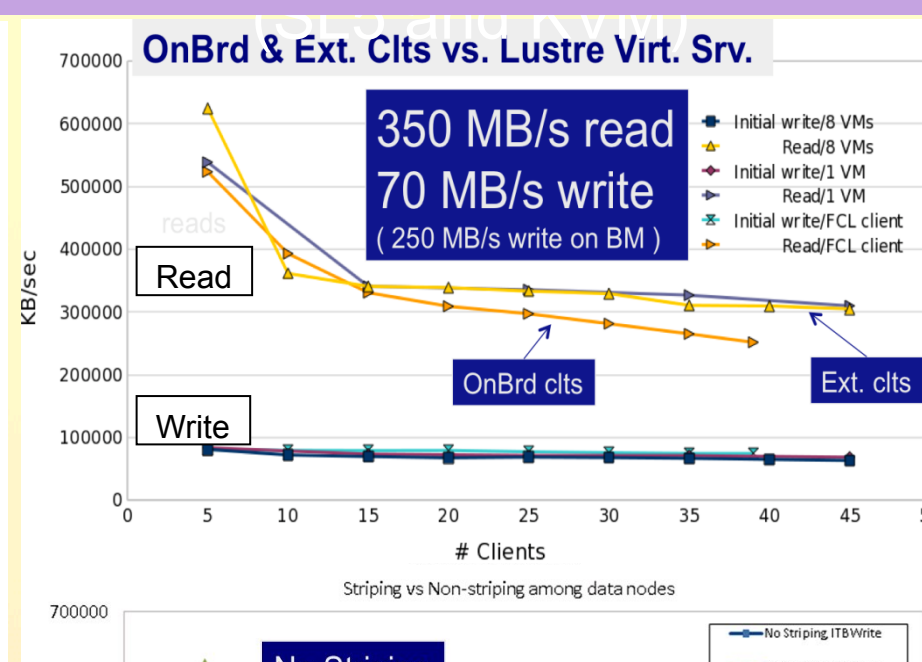
How well does Lustre perform with servers on Bare Metal vs. VM ?

Ext. Clts vs. Lustre BM Srv. (Baseline)



350 MB/s read
250 MB/s write

OnBrd & Ext. Clts vs. Lustre Virt. Srv.



350 MB/s read
70 MB/s write
(250 MB/s write on BM)

IOZone

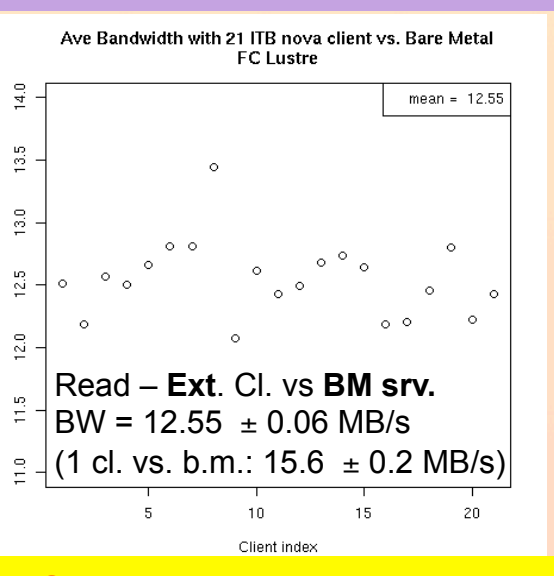
Read: same for Bare Metal and VM srv (w/ virtio net drv.)
Read: OnBrd clts 15% slower than Ext. clts (not significant)
Write: Bare Metal srv 3x faster than VM srv
Striping has a 5% effect on reading, none on writing.
No effect changing number of cores on Srv VM

Note: SL6 may have better write performance

Root Benchmark

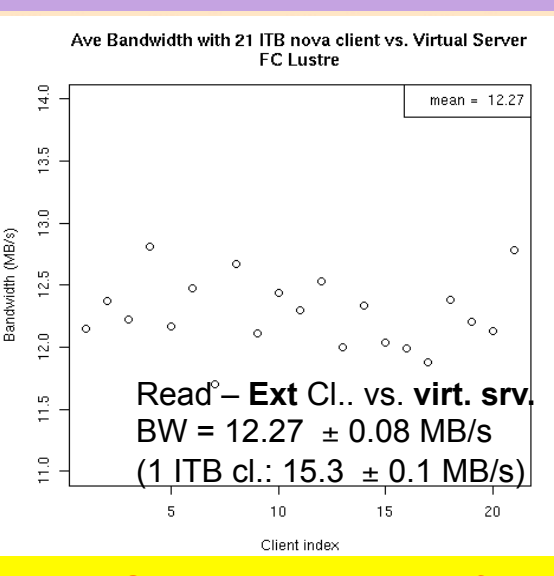
Read performance: how does Lustre Srv. on VM compare with Lustre Srv. on Bare Metal for External and On-Board clients?

Ave Bandwidth with 21 ITB nova client vs. Bare Metal FC Lustre



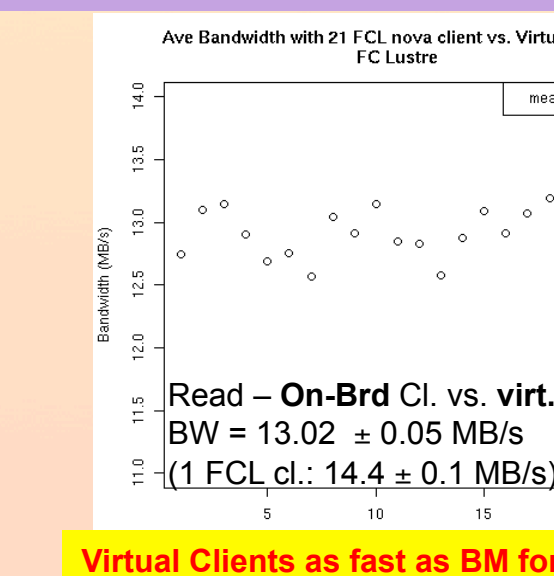
Read - Ext. Cl. vs BM srv.
BW = 12.55 ± 0.06 MB/s
(1 cl. vs. b.m.: 15.6 ± 0.2 MB/s)

Ave Bandwidth with 21 ITB nova client vs. Virtual Server FC Lustre



Read - Ext. Cl. vs. virt. srv.
BW = 12.27 ± 0.08 MB/s
(1 ITB cl.: 15.3 ± 0.1 MB/s)

Ave Bandwidth with 21 FCL nova client vs. Virtual Server FC Lustre



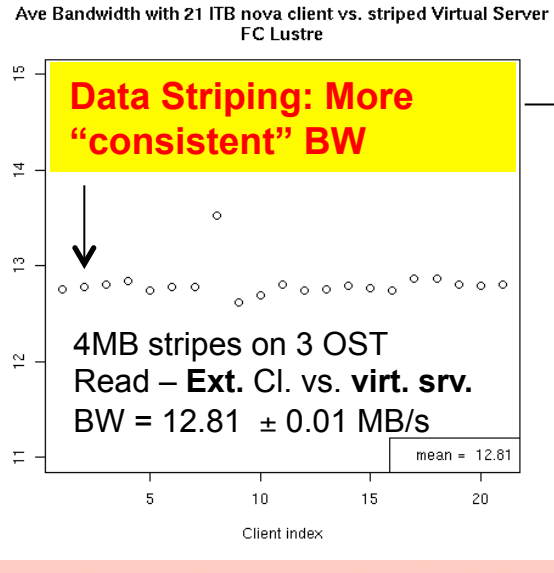
Read - On-Brd Cl. vs. virt. srv.
BW = 13.02 ± 0.05 MB/s
(1 FCL cl.: 14.4 ± 0.1 MB/s)

Non-Striped Bare Metal (BM) Server: baseline for read (ext. cl.)

Virtual Server is almost as fast as Bare Metal for read (ext. cl.)

Does Striping affect read BW for Ext. and On-Brd clients?

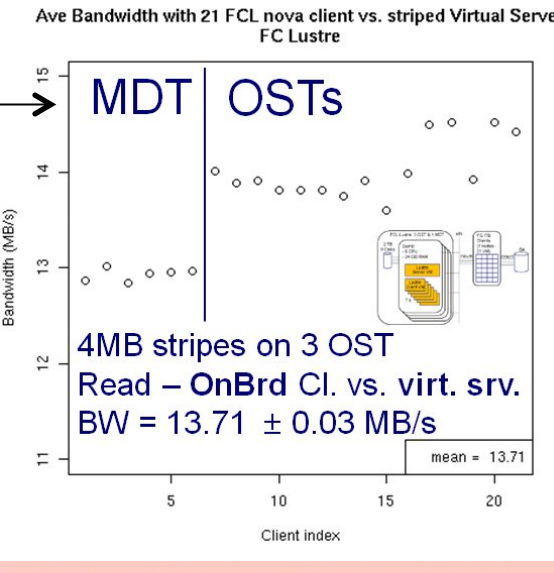
Ave Bandwidth with 21 ITB nova client vs. striped Virtual Server FC Lustre



Data Striping: More "consistent" BW

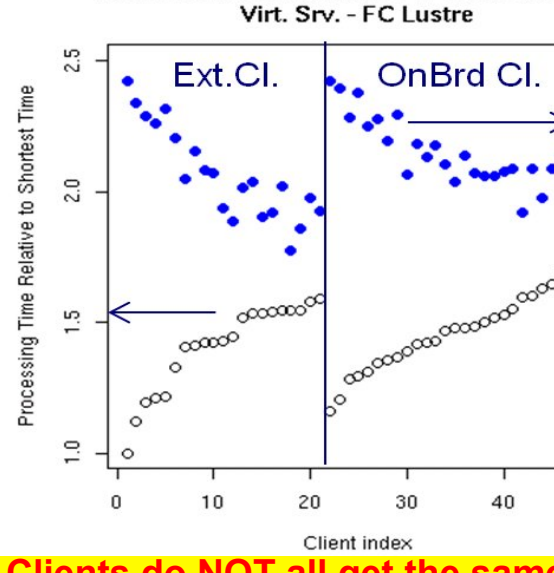
4MB stripes on 3 OST
Read - Ext. Cl. vs. virt. srv.
BW = 12.81 ± 0.01 MB/s

Ave Bandwidth with 21 FCL nova client vs. striped Virtual Server FC Lustre



4MB stripes on 3 OST
Read - OnBrd Cl. vs. virt. srv.
BW = 13.71 ± 0.03 MB/s

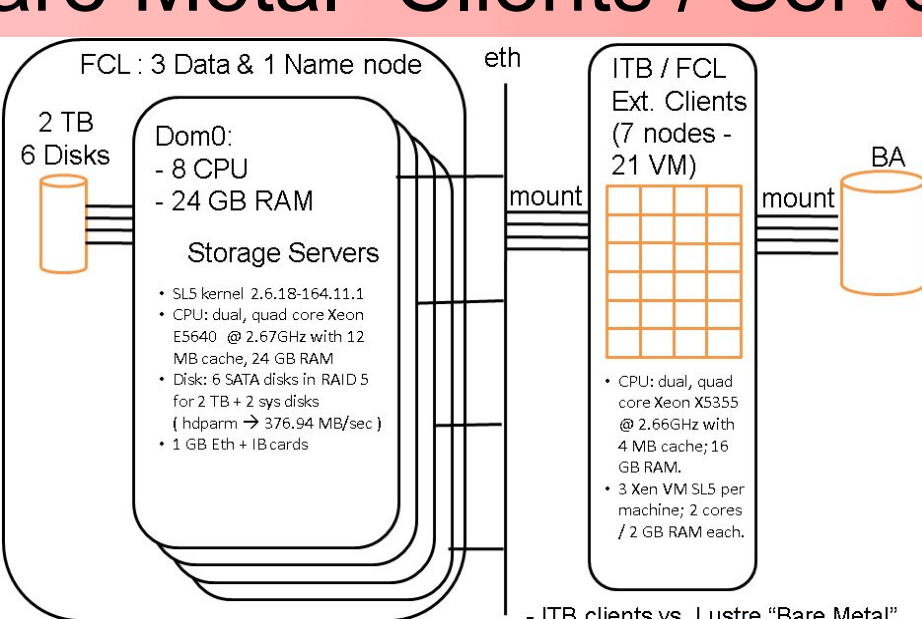
49 clts saturate the BW: is the distrib. fair?



Clients do NOT all get the same share of the bandwidth (within 20%).

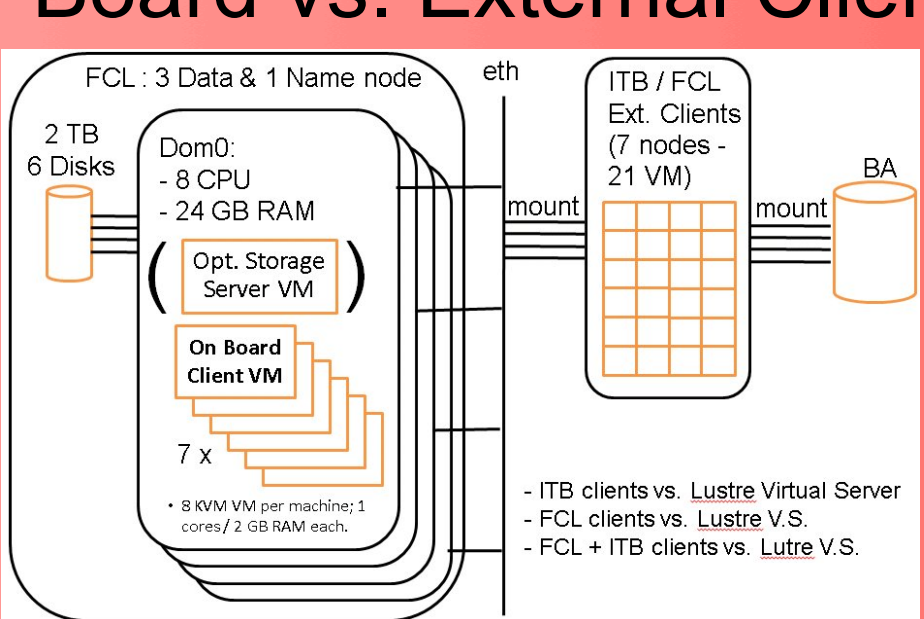
Storage Testbed

"Bare Metal" Clients / Servers



ITB clients vs. Lustre "Bare Metal"

On-Board vs. External Clients



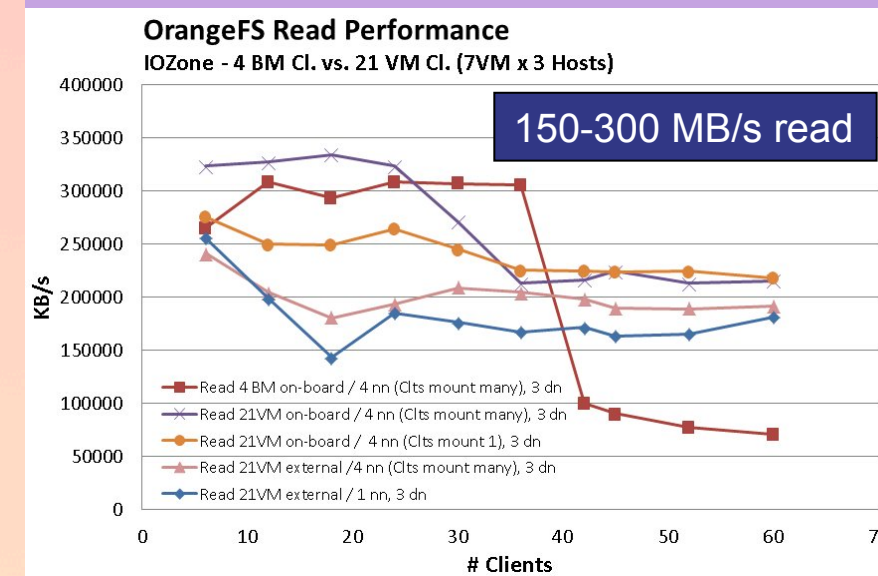
On-Brd Client VM run on the same host as the storage server.

OrangeFS

How well do VM clients perform vs. Bare Metal clients? Is there a difference for External vs. OnBoard clients? How does number of name nodes change performance?

OrangeFS Read Performance

IOZone - 4 BM Cl. vs. 21 VM Cl. (VPM x 3 Hosts)

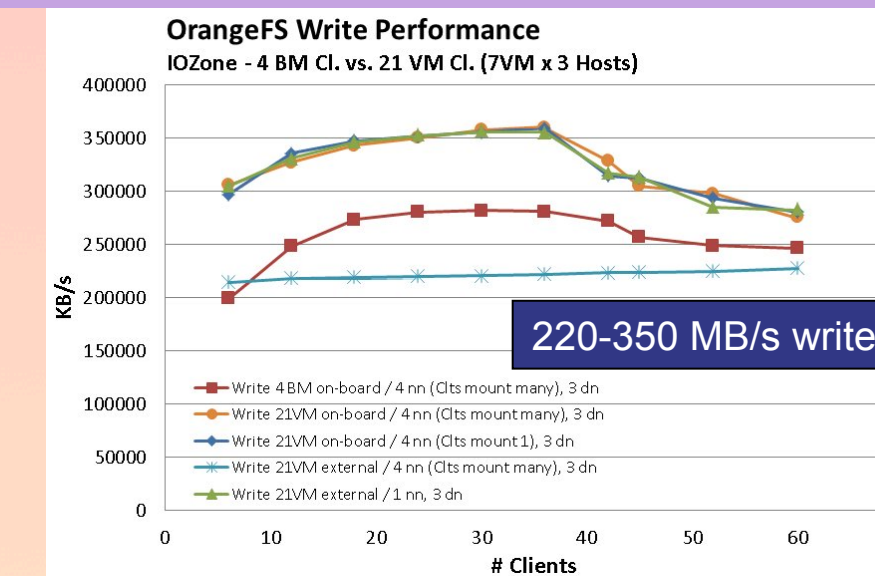


150-300 MB/s read

- On-Board Bare Metal clients read almost as fast as On-Board VM (faster config. w/ 4nn and 3 dn), but 50% slower than VM cl. for many processes; possibly too many procs for the OS to manage.
- On-Board VM Clients read 10%-60% faster than External VM clients.
- Using 4 name nodes improves read performance by 10%-60% as compared to 1 name node (different from write performance). Best performance when each name node serves a fraction of the clients.

OrangeFS Write Performance

IOZone - 4 BM Cl. vs. 21 VM Cl. (VPM x 3 Hosts)

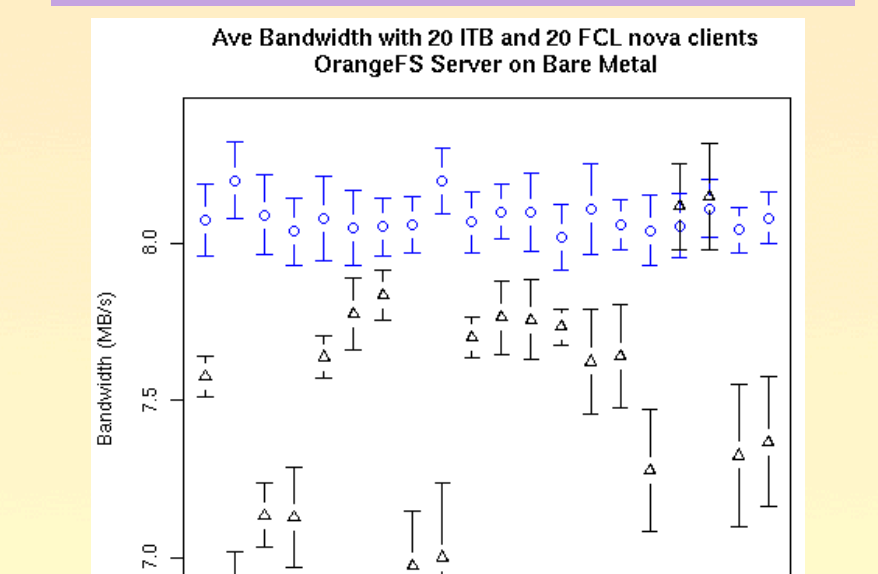


220-350 MB/s write

- On-Board Bare Metal clients write 80% slower than VM clients; possibly too many processes for the OS to manage.
- Write performance NOT consistent. On-Board VM clients generally have the same perf. as External VM clients. One reproducible 70% slower write meas. for External VM (4 name nodes when each nn serves a fraction of the cl.).
- Using 4 name nodes has 70% slower perf. for External VM (reproducible). Different from read performance.

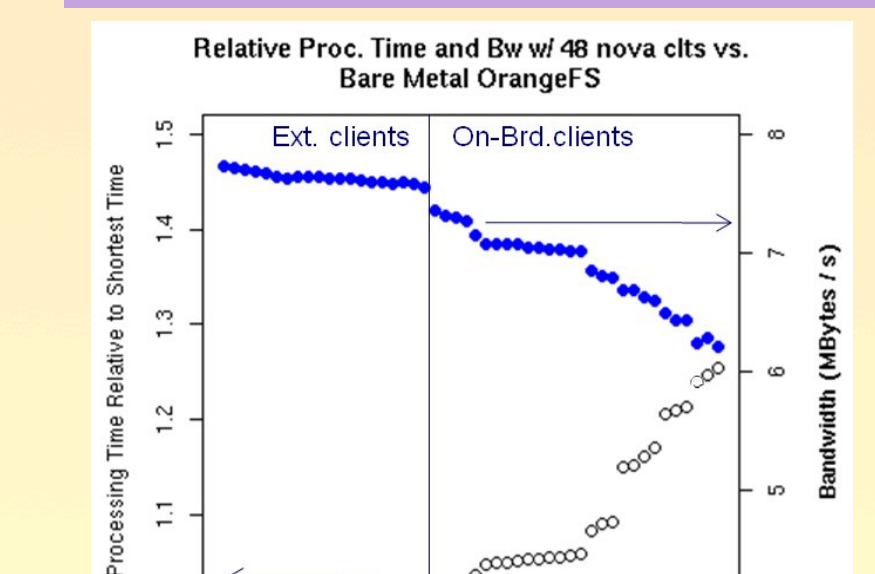
Root Benchmark

How does read BW vary for On-Brd vs. Ext. clnts?



External (ITB) clients read ~7% faster than on-board (FCL) clients (Same as Hadoop. Opposite from Lustre virt. Srv.).

49 clts (1 proc. / VM / core) saturate the BW to the srv. Is the distribution of the BW fair?



At saturation, on average External clients read ~10% faster than On-Board cl. (Same as Hadoop. Different from Lustre virt. Srv.).
External clients get the same share of the bw among themselves (within ~2%) (as Hadoop). On-Board clients have a larger spread (~20%) (as Lustre virtual Server.).